2017-01-19

Tom Rochette <tom.rochette@coreteks.org>

December 21, 2025 — 77e1b28a

0.1 Context

0.2 Learned in this study

- Undersampling and oversampling

0.3 Things to explore

- F2 score
- ROC/AUC
- SMOTE

1 Problems faced

- Working with np.array, dictionaries, lists, sets in order to uniquely identify array rows and index them

2 Overview

Based on questions from http://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/.

1) Can You Collect More Data?
   I certainly can, but it will not fix class imbalance in my case (data of the major class will always stay at the same proportion).

2) Try Changing Your Performance Metric
   Could be worth looking into, however it does not appear like it will be straightforward.

3) Try Resampling Your Dataset
   Undersampling and oversampling are two easy things that could be implemented. Furthermore, it wouldn't be very hard to make a generic function that receives training data and ensures proper class balance.

4) Try Generate Synthetic Samples
   Generating synthetic samples might be doable although kind of hard and most likely introducing false data.

5) Try Different Algorithms
   The article suggests using decision trees since they "often perform well on imbalanced data". It is worth giving it a try.

6) Try Penalized Models
   In this case, I'm not entirely sure how I'd apply it to my current model. This would require some work to understand the necessary changes. It might although be possible that simply giving a weight tensor to the `fit` method is enough to "penalize" during training.

7) Try a Different Perspective
It might be possible to look at the problem through different perspective, although that would require some ingenuity on my part (and I'd rather have the machine do that for me!).

8) Try Getting Creative
They suggest decomposing the larger class into smaller number of other classes, or basically trying to decompose the problem into sub-problems. That is one approach I had in mind, however I'm uncertain how decomposition would help since I assumed that the network would mostly do the decomposition itself through the internal/hidden layers. That may not be necessarily so.

3 See also

4 Keywords

- tensorflow imbalanced data
- keras class imbalance

5 References

- http://stackoverflow.com/questions/35049379/training-on-imbalanced-data-using-tensorflow
- https://www.reddit.com/r/MachineLearning/comments/12evgi/classification_when_80_of_my_training_set_is_of/
- http://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/
- http://contrib.scikit-learn.org/imbalanced-learn/generated/imblearn.over_sampling.SMOTE.html